A systematic review of natural language processing applied to radiology reports

RESEARCH

A Systematic Review of Natural Language Processing Applied to Radiology Reports

Arlene Casey[1*], Emma Davidson[2], Michael Poon[2], Hang Dong[3,4], Daniel Duma[1], Andreas Grivas[5], Claire Grover[5], Víctor Suárez-Paniagua[3,4], Richard Tobin[5], William Whiteley[2,6], Honghan Wu[4,7] and Beatrice Alex[1,8]

Correspondence:
rlene.Casey@ed.ac.uk
School of Literatures, Languages
nd Cultures (LLC), University of
dinburgh, Edinburgh, Scotland
ull list of author information is
vailable at the end of the article

Abstract

Background: Natural language processing (NLP) has a significant role in advancing healthcare and has been found to be key in extracting structured information from radiology reports. Understanding recent developments in NLP application to radiology is of significance but recent reviews on this are limited. This study systematically assesses and quantifies recent literature in NLP applied to radiology reports.

Methods: We conduct an automated literature search yielding 4,836 results using automated filtering, metadata enriching steps and citation search combined with manual review. Our analysis is based on 21 variables including radiology characteristics, NLP methodology, performance, study, and clinical application characteristics.

Results: We present a comprehensive analysis of the 164 publications retrieved with publications in 2019 almost triple those in 2015. Each publication is categorised into one of 6 clinical application categories. Deep learning use increases in the period but conventional machine learning approaches are still prevalent. Deep learning remains challenged when data is scarce and there is little evidence of adoption into clinical practice. Despite 17% of studies reporting greater than 0.85 F1 scores, it is hard to comparatively evaluate these approaches given that most of them use different datasets. Only 14 studies made their data and 15 their code available with 10 externally validating results.

Conclusions: Automated understanding of clinical narratives of the radiology reports has the potential to enhance the healthcare process and we show that research in this field continues to grow. Reproducibility and explainability of models are important if the domain is to move applications into clinical use. More could be done to share code enabling validation of methods on different institutional data and to reduce heterogeneity in reporting of study properties allowing inter-study comparisons. Our results have significance for researchers in the field providing a systematic synthesis of existing work to build on, identify gaps, opportunities for collaboration and avoid duplication.

Keywords: natural language processing; radiology; systematic review

Background

Medical imaging examinations interpreted by radiologists in the form of narrative reports are used to support and confirm diagnosis in clinical practice. Being able to accurately and quickly identify the information stored in radiologists' narratives has the potential to reduce workloads, support clinicians in their decision processes, triage patients to get urgent care or identify patients for research purposes. However, whilst these reports are generally considered more restricted in vocabulary than other electronic health records (EHR), e.g. clinical notes, it is still difficult to access this efficiently at scale [1]. This is due to the unstructured nature of these reports and Natural Language Processing (NLP) is key to obtaining structured information from radiology reports [2].

NLP applied to radiology reports is shown to be a growing field in earlier reviews [2, 3]. In recent years there has been an even more extensive growth in NLP research in general and in particular deep learning methods which is not seen in the earlier reviews. A more recent review of NLP applied to radiology-related research can be found but this focuses on one NLP technique only, deep learning models [4]. Our paper provides a more comprehensive review comparing and contrasting all NLP methodologies as they are applied to radiology.

It is of significance to understand and synthesise recent developments specific to NLP in the radiology research field as this will assist researchers to gain a broader understanding of the field, provide insight into methods and techniques supporting and promoting new developments in the field. Therefore, we carry out a systematic review of research output on NLP applications in radiology from 2015 onward, thus, allowing for a more up to date analysis of the area. An additional listing of our synthesis of publications detailing their clinical and technical categories can be found in Additional File 1 and per publication properties can be found in Additional File 2. Also different to the existing work, we look at both the clinical application areas NLP is being applied in and consider the trends in NLP methods. We describe and discuss study properties, e.g. data size, performance, annotation details, quantifying these in relation to both the clinical application areas and NLP methods. Having a more detailed understanding of these properties allows us to make recommendations for future NLP research applied to radiology datasets, supporting improvements and progress in this domain.

Related Work

Amongst pre-existing reviews in this area, [2] was the first that was both specific to NLP on radiology reports and systematic in methodology. Their literature search identified 67 studies published in the period up to October 2014. They examined the NLP methods used, summarised their performance and extracted the studies' clinical applications, which they assigned to five broad categories delineating their purpose. Since Pons et al.'s paper, several reviews have emerged with the broader remit of NLP applied to electronic health data, which includes radiology reports. [5] conducted a systematic review of NLP systems with a specific focus on coding free text into clinical terminologies and structured data capture. The systematic review by [6] specifically examined machine learning approaches to NLP (2015-2019) in more general clinical text data, and a further methodical review was carried out by [7] to synthesise literature on deep learning in clinical NLP (up to April 2019) although the did not follow the PRISMA guideline completely. With radiology reports as their particular focus, [3] published, the same year as Pons et al.'s review, an instructive narrative review outlining the fundamentals of NLP techniques applied in radiology. More recently, [4] published a systematic review focused on deep learning radiology-related research. They identified 10 relevant papers in their search (up to September 2019) and examined their deep learning models, comparing these with traditional NLP models and also considered their clinical applications but did not employ a specific categorisation. We build on this corpus of related work, and most specifically Pons et al.'s work. In our initial synthesis of clinical applications we adopt their application categories and further expand upon these to reflect the nature of subsequent literature captured in our work. Additionally, we quantify and compare properties of the studies reviewed and provide a series of recommendations for future NLP research applied to radiology datasets in order to promote improvements and progress in this domain.

Methods

Our methodology followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [8], and the protocol is registered on protocols.io.

Eligibility for Literature Inclusion and Search Strategy

We included studies using NLP on radiology reports of any imaging modality and anatomical region for NLP technical development, clinical support, or epidemiological research. Exclusion criteria included: (1) language not English; (2) wrong publication type (e.g., case reports, reviews, conference abstracts, comments, patents, or editorials) (2) published before 2015; (3) uses radiology images only (no NLP); (4) not radiology reports; (5) no NLP results; (6) year out of range; (7) duplicate, already in the list of publications retrieved; (8) not available in full text.

We used Publish or Perish [9], a citation retrieval and analysis software program, to search Google Scholar. Google Scholar has a similar coverage to other databases [10] and is easier to integrate into search pipelines. We conducted an initial pilot search following the process described here, but the search terms were too specific and restricted the number of publications. For example, we experimented with using specific terms used within medical imaging such at CT, MRI. Thirty-seven papers were found during the pilot search but the same papers also appeared in our final search. We use the following search query restricted to research articles published in English between January 2015 and October 2019. ("radiology" OR "radiologist") AND ("natural language" OR "text mining" OR "information extraction" OR "document classification" OR "word2vec") NOT patent. We automated the addition of publication metadata and applied filtering to remove irrelevant publications. These automated steps are described in Table 1 & Table 2.

Table 1 Metadata enriching steps undertaken for each publication

Metadata enriching steps
1. Match the paper with its DOI via the Crossref API [11]
2. If DOI matched, check Semantic Scholar for metadata/abstract [12]
3. If no DOI match and no abstract, search PubMed for abstract
4. Search arXiv [13] (for a pre-print)
5. If no PDF link, search Unpaywall for available open access versions [14]
6. If PDF but no separate abstract via Semantics Scholar/PubMed, extract abstract from the PDF

In addition to query search, another method to find papers is to conduct a citation search [15]. The citation search compiled a list of publications that cite the Pons et al. review and the articles cited in the Pons' review. To do this, we use a snowballing method [16] to follow the forward citation branch for each publication in this list, i.e. finding every article that cites the publications in our list. The branching factor here is large, so we filter at every stage and automatically add metadata. One hundred

Table 2 Automated filtering steps to remove irrelevant publications

Automated Filtering Steps
1. Document language is English
2. Word 'patent' in title or URL
3. Year of publication out of range (<2015)
4. The words 'review' or 'overview' in the title, 'this review' in the abstract
5. Image keywords in title or abstract with no NLP terminology in abstract
6. No radiology keywords in title or abstract
7. No NLP terminology in abstract

and seventy-one papers were identified as part of the snowball citation search and of these 84 were in the final 164 papers.

Manual Review of Literature

Four reviewers (three NLP researchers [AG,DD and HD] and one epidemiologist [MTCP]) independently screened all titles and abstracts with the Rayyan online platform and discussed disagreements. Fleiss' kappa [17] agreement between reviewers was 0.70, indicating substantial agreement [18]. After this screening process, each full-text article was reviewed by a team of eight (six NLP researchers and two epidemiologists) and double reviewed by a NLP researcher. We resolved any discrepancies by discussion in regular meetings.

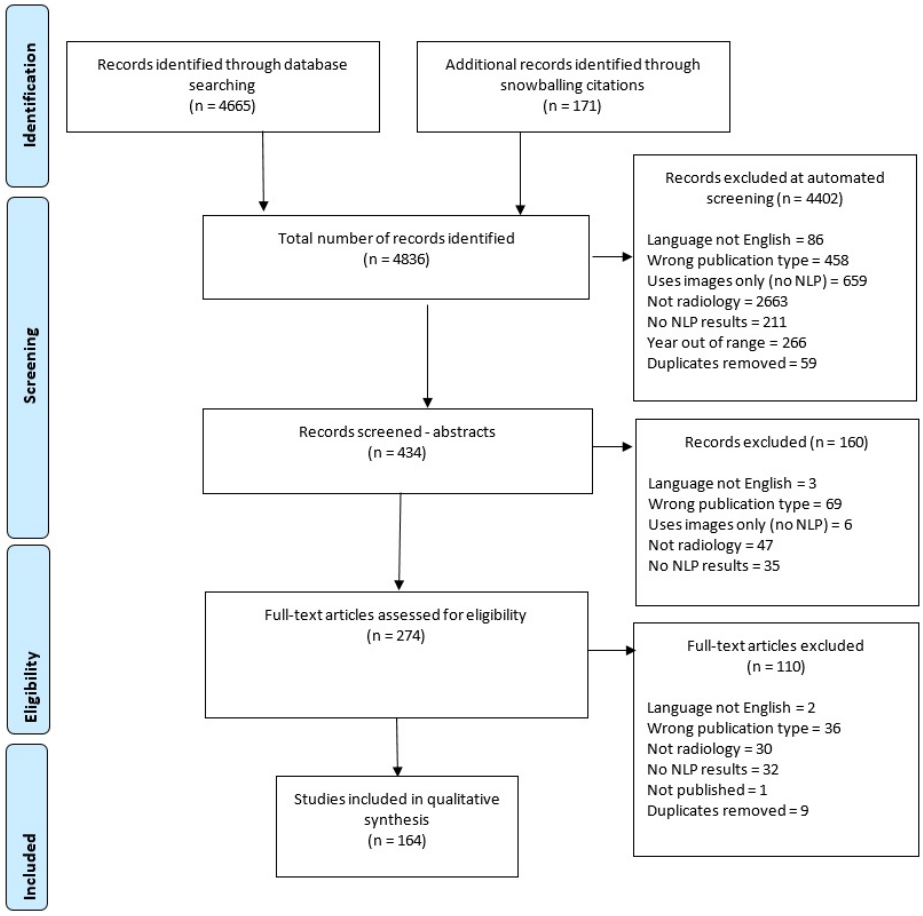
Data Extraction for Analysis

We extracted data on: primary clinical application and technical objective, data source(s), study period, radiology report language, anatomical region, imaging modality, disease area, dataset size, annotated set size, training/validation/test set size, external validation performed, domain expert used, number of annotators, inter-annotator agreement, NLP technique(s) used, best-reported results (recall, precision and F1 score), availability of dataset, and availability of code.

Results

The literature search yielded 4,836 possibly relevant publications from which our automated exclusion process removed 4,402, and during both our screening processes, 270 were removed, leaving 164 publications. See Figure 1 for details of exclusions at each step.

Figure 1 PRISMA diagram for search publication retrieval



General Characteristics

2015 and 2016 saw similar numbers of publications retrieved (22 and 21 respectively) with the volume increasing almost three-fold in 2019 (55), noting 2019 only covers 10 months (Figure 2). Imaging modality (Table 3) varied considerably and 46 studies used reports from multiple modalities. Of studies focusing on a single modality, the most featured were CT scans (38) followed by MRI (16), X-Ray (8), Mammogram (5) and Ultrasound (4). Forty-seven studies did not specifying scan modality. For the study samples (Table 4), 33 papers specified that they used consecutive patient images, 38 used non-consecutive image sampling and 93 did not clearly specify their sampling strategy. The anatomical regions for scans varied (Table 5) with mixed being the highest followed by Thorax and Head/neck. Disease categories are presented in Table 6 with the largest disease category being Oncology. The majority

of reports were in English (141) and a small number in other languages e.g., Chinese (5), Spanish (4), German (3) (Table 7). Additional file 2, CSV format, provides a breakdown of the information in Tables 3-7 per publication.

Table 3 Scan modality

Scan Modality	No. Studies
Multiple Modalities	46
MRI	16
CT	38
X-Ray	8
Mammogram	5
Ultrasound	4
Not specified	47
TOTAL	164

Table 4 Image sampling method

Sampling Method	No. Studies
Consecutive Images	33
Non-Consecutive Images	38
Not specified	93
TOTAL	164

Table 5 Anatomical region scanned

Anatomical Region	No. Studies
Mixed	43
Thorax	32
Head/Neck	25
Abdomen	15
Breast	15
Extremities	9
Spine	5
Other	1
Unspecified	19
TOTAL	164

Table 6 Disease category

Disease Category	No. Studies
Not specific disease related	40
Oncology	39
Various	20
Musculoskeletal	10
Cerebrovascular	13
Other	13
Respiratory	10
Trauma	7
Cardiovascular	6
Gastrointestinal	3
Hepatobiliary	2
Genitourinary	1
TOTAL	164

Table 7 Radiology report language

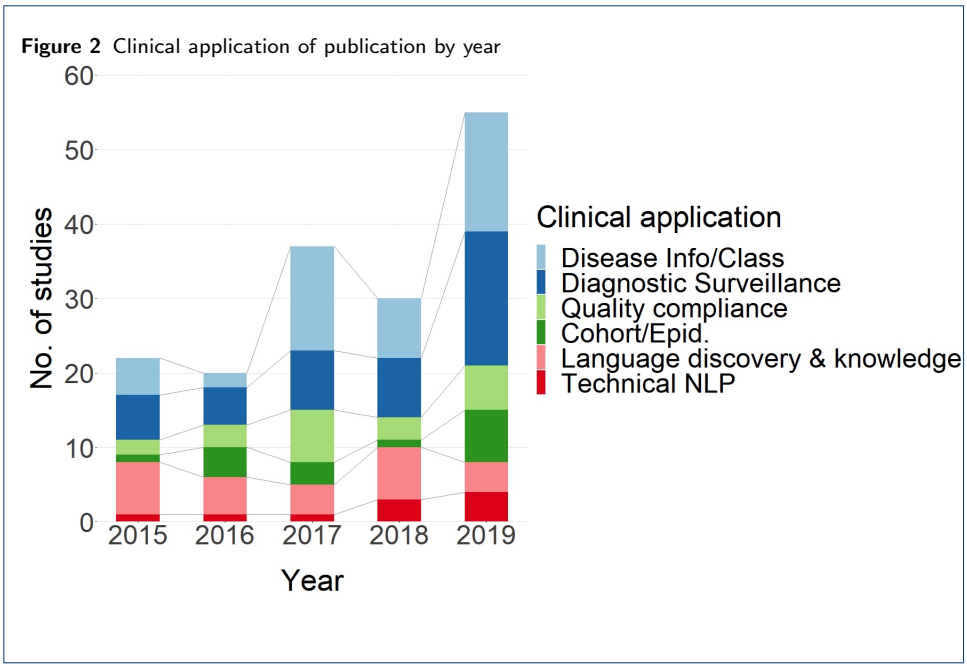
Report Language	No. Studies
English	141
Chinese	5
Spanish	4
German	3
Italian	2
French	2
Hebrew	1
Polish	1
Brazilian Portuguese	1
Unspecified	4
TOTAL	164

Table 8 Clinical application category by technical objective

Application Category	Information Extraction (n=73)	Report/ Sentence Classification (n=81)	Lexicon/ Ontology Discovery (n=9)	Clustering (n=1)
Disease Information & Classification	15	31 -	-	
Diagnostic Surveillance	28	17	-	-
Quality Compliance	5	15	-	-
Cohort-Epid.	6	10	-	-
Language Discovery & Knowledge	13	4	9	1
Technical NLP	6	4	-	-

Clinical Application Categories

In synthesis of the literature each publication was classified by the primary clinical purpose. Pons' work in 2016 categorised publications into 5 broad categories: Diagnostic Surveillance, Cohort Building for Epidemiological Studies, Query-based Case Retrieval, Quality Assessment of Radiological Practice and Clinical Support Services. We found some changes in this categorisation schema and our categorisation consisted of six categories: *Diagnostic Surveillance, Disease information and classification, Quality Compliance, Cohort/Epidemiology, Language Discovery and Knowledge Structure, Technical NLP*. The main difference is we found no evidence for a category of *Clinical Support Services* which described applications that had been integrated into the workflow to assist. Despite the increase in the number of publications, very few were in clinical use with more focus on the category of *Disease Information and Classification*. We describe each clinical application area in more detail below and where applicable how our categories differ from the earlier findings. A listing of all publications and their corresponding clinical application and technical category can be found in Additional File 1, MS Word format, and in Additional File 2 in CSV format. Table 8 shows the clinical application category by the technical classification and Figure 2 shows the breakdown of clinical application category by publication year. There were more publications in 2019 compared with 2015 for all categories except Language Discovery & Knowledge Structure, which fell by $\approx 25\%$ (Figure 2).



Diagnostic Surveillance

A large proportion of studies in this category focused on extracting disease information for patient or disease surveillance e.g. investigating tumour characteristics [19, 20]; changes over time [21] and worsening/progression or improvement/response to treatment [22, 23]; identifying correct anatomical labels [24]; organ measurements and temporality [25]. Studies also investigated pairing measurements between reports [26] and linking reports to monitoring changes through providing an integrated view of consecutive examinations [27]. Studies focused specifically on breast imaging findings investigating aspects, such as BI-RADS MRI descriptors (shape, size, margin) and final assessment categories (benign, malignant etc.) e.g., [28, 29, 30, 31, 32, 33]. Studies focused on tumour information e.g., for liver [34] and hepatocellular carcinoma (HPC) [35, 36] and one study on extracting information relevant for structuring subdural haematoma characteristics in reports [37].

Studies in this category also investigated incidental findings including on lung imaging [38, 39, 40], with [38] additionally extracting the nodule size; for trauma patients [41]; and looking for silent brain infarction and white matter disease [42]. Other studies focused on prioritising/triaging reports, detecting follow-up recommendations, and linking a follow-up exam to the initial recommendation report, or bio-surveillance of infectious conditions, such as invasive mould disease.

Disease Information and Classification

Disease Information and Classification publications use reports to identify information that may be aggregated according to classification systems. These publications focused solely on classifying a disease occurrence or extracting information about a disease with no focus on the overall clinical application. This category was not found in Pons' work. Methods considered a range of conditions including intracranial haemorrhage [43, 44], aneurysms [45], brain metastases [46], ischaemic stroke [47, 48], and several classified on types and severity of conditions e.g., [46, 49, 50, 51, 52]. Studies focused on breast imaging considered aspects such as predicting lesion malignancy from BI-RADS descriptors [53], breast cancer subtypes [54], and extracting or inferring BI-RADS categories, such as [55, 56]. Two studies focused on abdominal images and hepatocellular carcinoma (HCC) staging and CLIP scoring. Chest imaging reports were used to detect pulmonary embolism e.g., [57, 58, 59], bacterial pneumonia [60], and Lungs-RADS categories [61]. Functional imaging was also included, such as echocardiograms, extracting measurements to evaluate heart failure, including left ventricular ejection fractions (LVEF) [62]. Other studies investigated classification of fractures [63, 64] and abnormalities [65] and the prediction of ICD codes from imaging reports [66].

Language Discovery and Knowledge Structure

Language Discovery and Knowledge Structure publications investigate the structure of language in reports and how this might be optimised to facilitate decision support and communication. Pons et al. reported on applications of *Query-based retrieval* which has similarities to *Language Discovery and Knowledge Structure* but it is not the same. Their category contains studies that retrieve cases and conditions that are not predefined and in some instances could be used for research purposes or are motivated for educational purposes. Our category is broader and encompasses papers that investigated different aspects of language including variability, complexity simplification and normalising to support extraction and classification tasks.

Studies focus on exploring lexicon coverage and methods to support language simplification for patients looking at sources, such as the consumer health vocabulary [67] and the French lexical network (JDM) [68]. Other works studied the variability and complexity of report language comparing free-text and structured reports and

radiologists. Also investigated was how ontologies and lexicons could be combined with other NLP methods to represent knowledge that can support clinicians. This work included improving report reading efficiency [69]; finding similar reports [70]; normalising phrases to support classification and extraction tasks, such as entity recognition in Spanish reports [71]; imputing semantic classes for labelling [72], supporting search [73] or to discover semantic relations [74].

Quality and Compliance

Quality and Compliance publications use reports to assess the quality and safety of practice and reports similar to Pons' category. Works considered how patient indications for scans adhered to guidance e.g., [75, 76, 77, 78, 79, 80] or protocol selection [81, 82, 83, 84, 85] or the impact of guideline changes on practice, such as [86]. Also investigated was diagnostic utilisation and yield, based on clinicians or on patients, which can be useful for hospital planning and for clinicians to study their work patterns e.g.[87]. Other studies in this category looked at specific aspects of quality, such as, classification for long bone fractures to support quality improvement in paediatric medicine [88], automatic identification of reports that have critical findings for auditing purposes [89], deriving a query-based quality measure to compare structured and free-text report variability [90], and [91] who describe a method to fix errors in gender or laterality in a report.

Cohort and Epidemiology

This category is similar to Pons' earlier review but we treated the studies in this category differently attempting to differentiate which papers described methods for creating cohorts for research purposes, and those which also reported the outcomes of an epidemiological analysis. Ten studies use NLP to create specific cohorts for research purposes and six reported the performance of their tools. Out of these papers, the majority (n=8) created cohorts for specific medical conditions including fatty liver disease [92, 93] hepatocellular cancer [94], ureteric stones [95], vertebral fracture [96], traumatic brain injury [97, 98], and leptomeningeal disease secondary to metastatic breast cancer [99]. Five papers identified cohorts focused on particular radiology findings including ground glass opacities (GGO) [100], cerebral microbleeds (CMB) [101], pulmonary nodules [102], [103], changes in the spine correlated to back pain [1] and identifying radiological evidence of people having suffered a fall.

One paper focused on identifying abnormalities of specific anatomical regions of the ear within an audiology imaging database [104] and another paper aimed to create a cohort of people with any rare disease (within existing ontologies - Orphanet Rare Disease Ontology ORDO and Radiology Gamuts Ontology RGO). Lastly, one paper took a different approach of screening reports to create a cohort of people with contraindications for MRI, seeking to prevent iatrogenic events [105]. Amongst the epidemiology studies there were various analytical aims, but they primarily focused on estimating the prevalence or incidence of conditions or imaging findings and looking for associations of these conditions/findings with specific population demographics, associated factors or comorbidities. The focus of one study differed in that it applied NLP to healthcare evaluation, investigating the association of palliative care consultations and measures of high-quality end-of-life (EOL) care [99].

Technical NLP

This category is for publications that have a primary technical aim that is not focused on radiology report outcome, e.g. detecting negation in reports, spelling correction [106], fact checking [107, 108] methods for sample selection, crowd source annotation [109]. This category did not occur in Pons' earlier review.

NLP Methods in Use

NLP methods capture the different techniques an author applied broken down into rules, machine learning methods, deep learning, ontologies, lexicons and word embeddings. We discriminate machine learning from deep learning, using the former to represent traditional machine learning methods.

Over half of the studies only applied one type of NLP method and just over a quarter of the studies compared or combined methods in hybrid approaches. The remaining studies either used a bespoke proprietary system or focus on building ontologies or similarity measures (Figure 3). Rule-based method use remains almost constant across the period, whereas use of machine learning decreases and deep learning methods rises, from five publications in 2017 to twenty-four publications in 2019 (Figure 4).

A variety of machine classifier algorithms were used, with SVM and Logistic Regression being the most common (Table 9). Recurrent Neural Networks (RNN) variants were the most common type of deep learning architectures. RNN meth-

Figure 3 NLP method breakdown

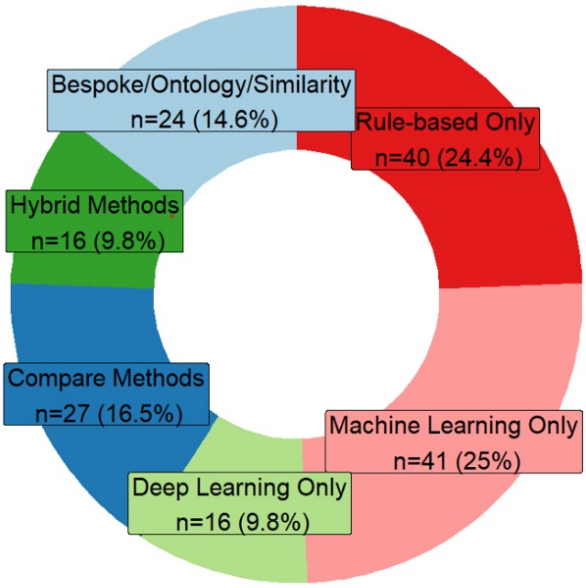
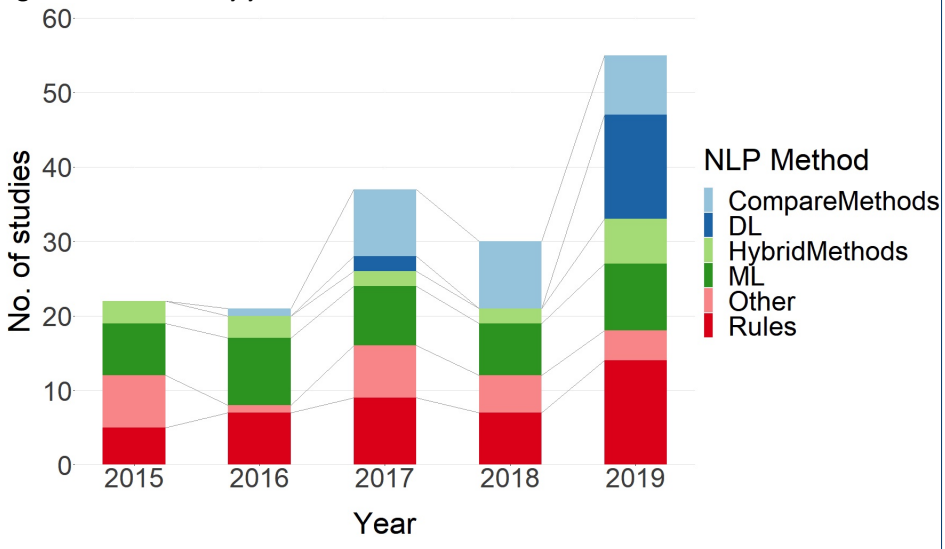


Figure 4 NLP method by year



ods were split between long short-term memory (LSTM) and bidirectional-LSTM (Bi-LSTM), bi-directional gated recurrent unit (Bi-GRU), and standard RNN approaches. Four of these studies additionally added a Conditional Random Field (CRF) for the final label generation step. Convolutional Neural Networks (CNN) were the second most common architecture explored. Eight studies additionally used

Table 9 Breakdown of NLP method

ML (n=74)	No studies	Deep Learning (n=36)	No studies
SVM	34	RNN variants	14
Logistic Regression	23	CNN	10
Random Forest	18	Other	5
Naïve Bayes	17	Compare CNN, RNN	4
Maximum Entropy	7	Combine CNN+RNN	3
Decision Trees	4		

an attention mechanism as part of their deep learning architecture. Other neural approaches included feed-forward neural networks, fully connected neural networks and a proprietary neural system IBM Watson [82] and Snorkel [110]. Several studies proposed combined architectures, such as [111, 31].

NLP Method Features

Most rule-based and machine classifying approaches used features based on bag-of-words, part-of-speech, term frequency, and phrases with only two studies alternatively using word embeddings. Three studies use feature engineering with deep learning rather than word embeddings. Thirty-three studies use domain-knowledge to support building features for their methods, such as developing lexicons or selecting terms and phrases. Comparison of embedding methods is difficult as many studies did not describe their embedding method. Of those that did, Word2Vec [112] was the most popular (n=19), followed by GLOVE embeddings [113] (n=6), FastText [114] (n=3), ELMo [115] (n=1) and BERT [116] (n=1). Ontologies or lexicon look-ups are used in 100 studies; however, even though publications increase over the period in real terms, 20% fewer studies employ the use of ontologies or lexicons in 2019 compared to 2015. The most widely used resources were UMLS [117] (n=15), Radlex [118] (n=20), SNOMED-CT [119] (n=14). Most studies used these as features for normalising words and phrases for classification, but this was mainly those using rule-based or machine learning classifiers with only six studies using ontologies as input to their deep learning architecture. Three of those investigated how existing ontologies can be combined with word embeddings to create domain-specific mappings, with authors pointing to this avoiding the need for large amounts of annotated data. Other approaches looked to extend existing medical resources using a frequent phrases approach, e.g. [120]. Works also used the derived

concepts and relations visualising these to support activities, such as report reading and report querying (e.g. [121, 122])

Annotation and Inter-Annotator Agreement

Eighty-nine studies used at least two annotators, 75 did not specify any annotation details, and only one study used a single annotator. Whilst 69 studies use a domain expert for annotation (a clinician or radiologist) only 56 studies report the inter-annotator agreement. Some studies mention annotation but do not report on agreement or annotators. Inter-annotator agreement values for Kappa range from 0.43 to perfect agreement at 1. Whilst most studies reported agreement by Cohen's Kappa [123] some reported precision, and percent agreement. Studies reported annotation data sizes differently, e.g., on the sentence or patient level. Studies also considered ground truth labels from coding schemes such as ICD or BI-RADS categories as annotated data. Of studies which detailed human annotation at the radiology report level, only 45 specified inter-annotator agreement and/or the number of annotators. Annotated report numbers for these studies varies with 15 papers having annotated less than 500, 12 having annotated between 500 and less than 1,000, 15 between 1,000 and less than 3,000, and 3 between 4,000 and 8,288 reports. Additional File 2 gives all annotation size information on a per publication basis in CSV format.

Data Sources and Availability

Only 14 studies reported that their data is available, and 15 studies reported that their code is available. Most studies sourced their data from medical institutions, a number of studies did not specify where their data was from, and some studies used publicly available datasets: MIMIC-III (n=5), MIMIC-II (n=1), MIMIC-CXR (n=1); Radcore (n=5) or STRIDE (n=2). Four studies used combined electronic health records such as clinical notes or pathology reports.

Reporting on total data size differed across studies with some not giving exact data sizes but percentages and others reporting numbers of sentences, reports, patients, or a mixture of these. Where an author was not clear on the type of data they were reporting on, or on the size, we marked this as unspecified. Thirteen studies did not report on total data size. Data size summaries for those reporting at the radiology report level is n=135 or 82.32% of the studies (Table 10). The biggest variation of data size by NLP Method is in studies that apply other methods or are rule-based.

Table 10 NLP Method by data size properties, minimum data size, maximum data size and median value, studies reporting in numbers of radiology reports

NLP Method	Min Size	Max Size	Median
Compare Methods	513	2,167,445	2,845
Hybrid Methods	40	34,926	918
Deep Learning (Only)	120	1,567,581	5,000
Machine Learning (Only)	101	2,977,739	2,531
Rules (Only)	31	10,000,000	8,000
Other	25	12,377,743	10,000

Table 11 Grouped data size and number of studies in each group, only for studies reporting in numbers of radiology reports

Data Size Group	No. Studies (%)
<200	9 (6.7)
200 <500	6 (4.4)
500 <1,000	18 (13.3)
1,000 <2,000	17 (12.6)
2,000 <5,000	17 (12.6)
5,000 <10,000	12 (8.9)
10,000+	53 (39.3)
Unspecified	3 (2.2)

Table 12 Studies reporting on total data size used and details on training set size, validation set size, test set size and annotation set size

Dataset Type	No. of Studies	Comments
Total Dataset Size	151	5
Training Set Size	129	
Validation Set Size	52	27 report size, 25 report k-fold cross validation
Test Set Size	81	
Annotation Set Size	97	

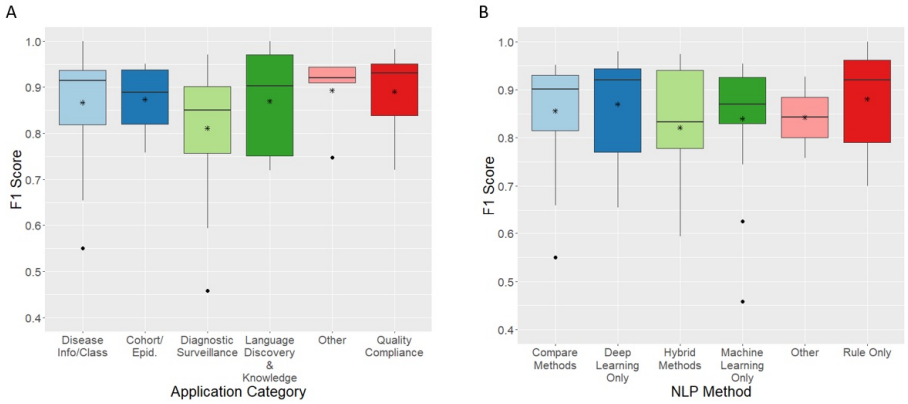
Machine learning also varies in size; however, the median value is lower compared to rule-based methods. The median value for deep learning is considerably higher at 5,000 reports compared to machine learning or those that compare or create hybrid methods. Of the studies reporting on radiology reports numbers, 39.3% used over 10,000 reports and this increases to over 48% using more than 5,000 reports. However, a small number of studies, 14%, are using comparatively low numbers of radiology reports, less than 500 (Table 11).

NLP Performance and Evaluation Measures

Performance metrics applied for evaluation of methods vary widely with authors using precision (positive predictive value (PPV)), recall (sensitivity), specificity, the area under the curve (AUC) or accuracy. We observed a wide variety in evaluation methodology employed concerning test or validation datasets. Different approaches were taken in generating splits for testing and validation, including k-fold cross-validation. Table 12 gives a summary of the number of studies reporting about total data size and splits across train, validation, test, and annotation. This table is for all data types, i.e., reports, sentences, patients or mixed. Eighty-two studies reported on both training and test data splits, of which only 38 studies included a validation set. Only 10 studies validated their algorithm using an external dataset from another institution, another modality, or a different patient population. Additional File 2 gives all data size information on a per publication basis in CSV format.

The most widely used metrics for reporting performance were precision (PPV) and recall (sensitivity) reported in 47% of studies. However, even though many studies compared methods and reported on the top-performing method, very few studies carried out significance testing on these comparisons. Issues of heterogeneity make it difficult and unrealistic to compare performance between methods applied, hence, we use summary measures as a broad overview (Figure 5). Performance reported varies, but both the mean and median values for the F1 score appear higher for methods using rule-based only or deep learning only methods. Whilst differences are less discernible between F1 scores for application areas, *Diagnostic Surveillance* looks on average lower than other categories.

Figure 5 Application Category and NLP Method, Mean and Median Summaries. Mean value is indicated by a vertical bar, the box shows error bars and the asterisk is the median value.



Discussion and Future Directions

Our work shows there has been a considerable increase in the number of publications using NLP on radiology reports over the recent time period. Compared to 67 publications retrieved in the earlier review of [2], we retrieved 164 publications. In this section we discuss and offer some insight into the observations and trends of how NLP is being applied to radiology and make some recommendations that may benefit the field going forward.

Clinical Applications and NLP Methods in Radiology

The clinical applications of the publications is similar to the earlier review of Pons et al. but whilst we observe an increase in research output we also highlight that there appears to be even less focus on clinical application compared to their review. Like many other fields applying NLP the use of deep learning has increased, with RNN architectures being the most popular. This is also observed in a review of NLP in clinical text[7]. However, although deep learning use increases, rules and traditional machine classifiers are still prevalent and often used as baselines to compare deep learning architectures against. One reason for traditional methods remaining popular is their interpretability compared to deep learning models. Understanding the features that drive a model prediction can support decision-making in the clinical domain but the complex layers of non-linear data transformations deep learning is composed of does not easily support transparency [124]. This may also help ex-

plain why in synthesis of the literature we observed less focus on discussing clinical application and more emphasis on disease classification or information task only. Advances in interpretability of deep learning models are critical to its adoption in clinical practice.

Other challenges exist for deep learning such as only having access to small or imbalanced datasets. Chen et al. [125] review deep learning methods within healthcare and point to these challenges resulting in poor performance but that these same datasets can perform well with traditional machine learning methods. We found several studies highlight this and when data is scarce or datasets imbalanced, they introduced hybrid approaches of rules and deep learning to improve performance, particularly in the *Diagnostic Surveillance* category. Yang et al. [126] observed rules performing better for some entity types, such as time and size, which are proportionally lower than some of the other entities in their train and test sets; hence they combine a bidirectional-LSTM and CRF with rules for entity recognition. Peng et al. [19] comment that combining rules and the neural architecture complement each other, with deep learning being more balanced between precision and recall, but the rule-based method having higher precision and lower recall. The authors reason that this provides better performance as rules can capture rare disease cases, particularly when multi-class labelling is needed, whilst deep learning architectures perform worse in instances with fewer data points.

In addition to its need for large-scale data, deep learning can be computationally costly. The use of pre-trained models and embeddings may alleviate some of this burden. Pre-trained models often only require fine-tuning, which can reduce computation cost. Language comprehension pre-learned from other tasks can then be inherited from the parent models, meaning fewer domain-specific labelled examples may be needed [127]. This use of pre-trained information also supports generalisability, e.g., [58] show that their model trained on one dataset can generalise to other institutional datasets.

Embedding use has increased which is expected with the application of deep learning approaches but many rule-based and machine classifiers continue to use traditional count-based features, e.g., bag-of-words and n-grams. Recent evidence [128] suggests that the trend to continue to use feature engineering with traditional

machine learning methods does produce better performance in radiology reports than using domain-specific word embeddings.

Banerjee et al. [44] found that there was not much difference between a uni-gram approach and a Word2vec embedding, hypothesising this was due to their narrow domain, intracranial haemorrhage. However, the NLP research field has seen a move towards bi-directional encoder representations from transformers (BERT) based embedding models not reflected in our analysis, with only one study using BERT generated embeddings [46]. Embeddings from BERT are thought to be superior as they can deliver better contextual representations and result in improved task performance. Whilst more publications since our review period have used BERT based embeddings with radiology reports e.g. [127, 129] not all outperform traditional methods [130]. Recent evidence shows that embeddings generated by BERT fail to show a generalisable understanding of negation [131], an essential factor in interpreting radiology reports effectively. Specialised BERT models have been introduced such as ClinicalBERT [132] or BlueBERT [129]. BlueBERT has been shown to outperform ClinicalBERT when considering chest radiology [133] but more exploration of the performance gains versus the benefits of generalisability are needed for radiology text.

All NLP models have in common that they need large amounts of labelled data for model training [134]. Several studies [135, 136, 137] explored combining word embeddings and ontologies to create domain-specific mappings, and they suggest this can avoid a need for large amounts of annotated data. Additionally, [135, 136] highlight that such combinations could boost coverage and performance compared to more conventional techniques for concept normalisation.

The number of publications using medical lexical knowledge resources is still relatively low, even though a recent trend in the general NLP field is to enhance deep learning with external knowledge [138]. This was also observed by [7], where only 18% of the deep learning studies in their review utilised knowledge resources. Although pre-training supports learning previously known facts it could introduce unwanted bias, hindering performance. The inclusion of domain expertise through resources such as medical lexical knowledge may help reduce this unwanted bias [7]. Exploration of how this domain expertise can be incorporated with deep learning

architectures in future could improve the performance when having access to less labelled data.

Task Knowledge

Knowledge about the disease area of interest and how aspects of this disease are linguistically expressed is useful and could promote better performing solutions. Whilst [139] find high variability between radiologists, with metric values (e.g. number of syntactic, clinical terms based on ontology mapping) being significantly greater on free-text than structured reports, [140] who look specifically at anatomical areas find less evidence for variability. Zech et al. [141] suggest that the highly specialised nature of each imaging modality creates different sub-languages and the ability to discover these labels (i.e. disease mentions) reflects the consistency with which labels are referred to. For example, edema is referred to very consistently whereas other labels are not, such as infarction/ischaemic. Understanding the language and the context of entity mentions could help promote novel ideas on how to solve problems more effectively. For example, [35] discuss how the accuracy of predicting malignancy is affected by cues being outside their window of consideration and [142] observe problems of co-reference resolution within a report due to long-range dependencies. Both these studies use traditional NLP approaches, but we observed novel neural architectures being proposed to improve performance in similar tasks specifically capturing long-range context and dependency learning, e.g., [111, 31]. This understanding requires close cooperation of healthcare professionals and data scientists, which is different to some other fields where more disconnection is present [125].

Study Heterogeneity, a Need for Reporting Standards

Most studies reviewed could be described as a proof-of-concept and not trialled in a clinical setting. Pons et al. [2] hypothesised that a lack of clinical application may stem from uncertainty around minimal performance requirements hampering implementations, evidence-based practice requiring justification and transparency of decisions, and the inability to be able to compare to human performance as the human agreement is often an unknown. These hypotheses are still valid, and we see little evidence that these problems are solved.

Human annotation is generally considered the gold standard at measuring human performance, and whilst many studies reported that they used annotated data, overall, reporting was inconsistent. Steps were undertaken to measure inter-annotator agreement (IAA), but in many studies, this was not directly comparable to the evaluation undertaken of the NLP methods. The size of the data being used to draw experimental conclusions from is important and accurate reporting of these measures is essential to ensure reproducibility and comparison in further studies. Reporting on the training, test and validation splits was varied with some studies not giving details and not using held-out validation sets.

Most studies use retrospective data from single institutions but this can lead to a model over-fitting and, thus, not generalising well when applied in a new setting. Overcoming the problem of data availability is challenging due to privacy and ethics concerns, but essential to ensure that performance of models can be investigated across institutions, modalities, and methods. Availability of data would allow for agreed benchmarks to be developed within the field that algorithm improvements can be measured upon. External validation of applied methods was extremely low, although, this is likely due to the availability of external datasets. Making code available would enable researchers to report how external systems perform on their data. However, only 15 studies reported that their code is available. To be able to compare systems there is a need for common datasets to be available to benchmark and compare systems against.

Whilst reported figures in precision and recall generally look high more evidence is needed for accurate comparison to human performance. A wide variety of performance measures were used, with some studies only reporting one measure, e.g., accuracy or F1 scores, with these likely representing the best performance obtained. Individual studies are often not directly comparable for such measures, but nonethe-less clarity and consistency in reporting is desirable. Many studies making model comparisons did not carry out any significance testing for these comparisons.

Progressing NLP in radiology

The value of NLP applied to radiology is clear in that it can support areas such as clinicians in their decision making and reducing workload, add value in terms of automated coding of data, finding missed diagnosis for triage or monitoring quality.

However, in recent years labelling disease phenotypes or extracting disease information in reports has been a focus rather than real-world clinical application of NLP within radiology. We believe this is mainly due to the difficulties in accessing data for research purposes. More support is needed to bring clinicians and NLP experts together to promote innovative thinking about how such work can benefit and be trialled in the clinical environment. The challenges in doing so are significant because of the need to work within safe environments to protect patient privacy. In terms of NLP methods, we observe that the general trends of NLP are applied within this research area, but we would emphasise as NLP moves more to deep learning it is particularly important in healthcare to think about how these methods can satisfy explainability. Explainability in artificial intelligence and NLP has become a hot topic in general but it is now also being addressed in the healthcare sector [143, 144]. Methodology used is also impacted by data availability with uncommon diseases often being hard to predict with deep learning as data is scarce. If the practical and methodological challenges on data access, privacy and less data demanding approaches can be met there is much potential to increase the value of NLP within radiology. The sharing of tools, practice, and expertise could also ease the real-world application of NLP within radiology.

To help move the field forward, enable more inter-study comparisons, and increase study reproducibility we make the following recommendations for research studies:

1. Clarity in reporting study properties is required: (a) Data characteristics including size and the type of dataset should be detailed, e.g., the number of reports, sentences, patients, and if patients how many reports per patient. The training, test and validation data split should be evident, as should the source of the data. (b) Annotation characteristics including the methodology to develop the annotation should be reported, e.g., annotation set size, annotator details, how many, expertise. (c) Performance metrics should include a range of metrics: precision, recall, F1, accuracy and not just one overall value.
2. Significance testing should be carried out when a comparison between methods is made.
3. Data and code availability are encouraged. While making data available will often be challenging due to privacy concerns, researchers should make code

available to enable inter-study comparisons and external validation of methods.

4 Common datasets should be used to benchmark and compare systems.

Limitations of Study

Publication search is subject to bias in search methods and it is likely that our search strategy did inevitably miss some publications. Whilst trying to be precise and objective during our review process some of the data collected and categorising publications into categories was difficult to agree on and was subjective. For example, many of the publications could have belonged to more than one category. One of the reasons for this was how diverse in structure the content was which was in some ways reflected by the different domains papers were published in. It is also possible that certain keywords were missed in recording data elements due to the reviewers own biases and research experience.

Conclusions

This paper presents an systematic review of publications using NLP on radiology reports during the period 2015 to October 2019. We show there has been substantial growth in the field particularly in researchers using deep learning methods. Whilst deep learning use has increased, as seen in NLP research in general, it faces challenges of lower performance when data is scarce or when labelled data is unavailable, and is not widely used in clinical practice perhaps due to the difficulties in interpretability of such models. Traditional machine learning and rule-based methods are, therefore, still widely in use. Exploration of domain expertise such as medial lexical knowledge must be explored further to enhance performance when data is scarce. The clinical domain faces challenges due to privacy and ethics in sharing data but overcoming this would enable development of benchmarks to measure algorithm performance and test model robustness across institutions. Common agreed datasets to compare performance of tools against would help support the community in inter-study comparisons and validation of systems. The work we present here has the potential to inform researchers about applications of NLP to radiology and to lead to more reliable and responsible research in the domain.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files]

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Alan Turing Institute, MRC, HDR-UK and the Chief Scientist Office. B.A.,A.C,D.D.,A.G. and C.G. have been supported by the Alan Turing Institute via Turing Fellowships (B.A,C.G.) and Turing project funding (ESPRC grant EP/N510129/1). A.G. was also funded by a MRC Mental Health Data Pathfinder Award (MRC-MCPC17209). H.W. is MRC/Rutherford Fellow HRD UK (MR/S004149/1). H.D. is supported by HDR UK National Phemomics Resource Project. V.S-P. is supported by the HDR UK National Text Analytics Implementation Project. W.W. is supported by a Scottish Senior Clinical Fellowship (CAF/17/01).

Authors' contributions

B.A, W.W. and H.W. conceptualised this study. D.D. carried out the search including automated filtering and designing meta-enriching steps. BA, AG, CG and RT advised on the automatic data collection method devised by DD. M.T.C.P, A.G., H.D. and D.D carried out the first stage review and A.C., E.D., V.S-P, M.T.C.P, A.G., H.D., B.A. and D.D. carried out the second-stage review. A.C. synthesised the data and wrote the main manuscript with contributions from all authors. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Abbreviations

NLP - natural language processing
e.g. - example
ICD - international classification of diseases
BI-RADS - Breast Imaging-Reporting and Data System
IAA - inter-annotator agreement
No. - number
UMLS - unified medical language system
ELMo - embeddings from Language Models
BERT - bidirectional encoder representations from transformers
SVM - support vector machine
CNN - convolutional neural network
LSTM - long short-term memory
Bi-LSTM - bi-directional long short-term memory
Bi-GRU - bi-directional gated recurrent unit
CRF - conditional random field
GLOVE - Global Vectors for Word Representation

Author details

[1]School of Literatures, Languages and Cultures (LLC), University of Edinburgh, Edinburgh, Scotland. [2]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, Scotland. [3]Centre for Medical Informatics, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland. [4]Health Data Research UK , London, U.K.. [5]Institute for Language, Cognition and Computation, School of informatics,

University of Edinburgh, Edinburgh, Scotland. [6]Nuffield Department of Population Health, University of Oxford, Oxford, U.K.. [7]Institute of Health Informatics, University College London, London, U.K.. [8]Edinburgh Futures Institute, University of Edinburgh, Edinburgh, Scotland.

References

1. Bates, J., Fodeh, S.J., Brandt, C.A., Womack, J.A.: Classification of radiology reports for falls in an HIV study cohort. Journal of the American Medical Informatics Association **23**(e1), 113–117 (2016). doi:10.1093/jamia/ocv155. Accessed 2020-10-30
2. Pons, E., Braun, L.M.M., Hunink, M.G.M., Kors, J.A.: Natural Language Processing in Radiology: A Systematic Review. Radiology **279**(2), 329–343 (2016). doi:10.1148/radiol.16142770. Accessed 2020-10-30
3. Cai, T., Giannopoulos, A.A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K.K., Rybicki, F.J., Mitsouras, D.: Natural Language Processing Technologies in Radiology Research and Clinical Applications. RadioGraphics **36**(1), 176–191 (2016). doi:10.1148/rg.2016150080. Accessed 2020-10-30
4. Sorin, V., Barash, Y., Konen, E., Klang, E.: Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. Journal of the American College of Radiology : JACR **17**(5), 639–648 (2020). doi:10.1016/j.jacr.2019.12.026
5. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., Botsis, T.: Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics **73**, 14–29 (2017). doi:10.1016/j.jbi.2017.07.012
6. Spasic, I., Nenadic, G.: Clinical Text Data in Machine Learning: Systematic Review. JMIR medical informatics **8**(3), 17984 (2020). doi:10.2196/17984
7. Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., Xu, H.: Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association: JAMIA **27**(3), 457–470 (2020). doi:10.1093/jamia/ocz200
8. Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A.: Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews **4**(1), 1 (2015). doi:10.1186/2046-4053-4-1. Accessed 2020-11-05
9. Harzing A. W.: Publish or Perish, (2007). Available from https://harzing.com/resources/publish-or-perish
10. Gehanno, J.-F., Rollin, L., Darmoni, S.: Is the coverage of google scholar enough to be used alone for systematic reviews. BMC Medical Informatics and Decision Making **13**, 7 (2013). doi:10.1186/1472-6947-13-7. Accessed 2020-10-31
11. Wilkinson, L.J.: REST API. Publication Title: Crossref Type: website. https://www.crossref.org/education/retrieve-metadata/rest-api/ Accessed 2020-01-26
12. for AI, A.I.: Semantic Scholar \textbar AI-Powered Research Tool. https://api.semanticscholar.org/ Accessed 2021-01-26
13. University, C.: arXiv.org e-Print archive. https://arxiv.org/ Accessed 2021-01-26
14. Bearden, E.: LibGuides: Unpaywall: Home. https://library.lasalle.edu/c.php?g=982604&p=7105436 Accessed 2021-01-26
15. Briscoe, S., Bethel, A., Rogers, M.: Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study. Research Synthesis Methods **11**(2), 169–180 (2020). doi:10.1002/jrsm.1355. Accessed 2020-10-31
16. Wohlin, C.: Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. EASE '14. Association for Computing Machinery, New York, NY, USA (2014). doi:10.1145/2601248.2601268. event-place: London, England, United Kingdom. https://doi.org/10.1145/2601248.2601268
17. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971). doi:10.1037/h0031619
18. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1), 159–174 (1977). doi:10.2307/2529310. Accessed 2020-10-31
19. Peng, Y., Yan, K., Sandfort, V., Summers, R.M., Lu, Z.: A self-attention based deep learning method for lesion attribute detection from CT reports. In: 2019 IEEE International Conference on Healthcare Informatics

(ICHI), pp. 1–5. IEEE Computer Society, Xi'an, China (2019). doi:10.1109/ICHI.2019.8904668

20. Bozkurt, S., Alkim, E., Banerjee, I., Rubin, D.L.: Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. Journal of Digital Imaging **32**(4), 544–553 (2019). doi:10.1007/s10278-019-00237-9. Accessed 2020-10-30
21. Hassanpour, S., Bay, G., Langlotz, C.P.: Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. Journal of Digital Imaging **30**(3), 314–322 (2017). doi:10.1007/s10278-016-9931-8. Accessed 2020-10-30
22. Kehl, K.L., Elmarakeby, H., Nishino, M., Van Allen, E.M., Lepisto, E.M., Hassett, M.J., Johnson, B.E., Schrag, D.: Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. JAMA Oncology **5**(10), 1421–1429 (2019). doi:10.1001/jamaoncol.2019.1800. Accessed 2020-10-30
23. Chen, P.-H., Zafar, H., Galperin-Aizenberg, M., Cook, T.: Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. Journal of Digital Imaging **31**(2), 178–184 (2018). doi:10.1007/s10278-017-0027-x. Accessed 2020-10-30
24. Cotik, V., Rodríguez, H., Vivaldi, J.: Spanish Named Entity Recognition in the Biomedical Domain. In: Lossio-Ventura, J.A., Muñante, D., Alatrista-Salas, H. (eds.) Information Management and Big Data. Communications in Computer and Information Science, vol. 898, pp. 233–248. Springer, Lima, Peru (2018). doi:10.1007/978-3-030-11680-4-23
25. Sevenster, M., Buurman, J., Liu, P., Peters, J.F., Chang, P.J.: Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. Applied Clinical Informatics **06**(3), 600–610 (2015). doi:10.4338/ACI-2014-11-RA-0110. Accessed 2020-10-30
26. Sevenster, M., Bozeman, J., Cowhy, A., Trost, W.: A natural language processing pipeline for pairing measurements uniquely across free-text CT reports. Journal of Biomedical Informatics **53**, 36–48 (2015). doi:10.1016/j.jbi.2014.08.015. Accessed 2020-10-30
27. Oberkampf, H., Zillner, S., Overton, J.A., Bauer, B., Cavallaro, A., Uder, M., Hammon, M.: Semantic representation of reported measurements in radiology. BMC Medical Informatics and Decision Making **16**(1), 5 (2016). doi:10.1186/s12911-016-0248-9. Accessed 2020-10-30
28. Liu, Y., Zhu, L.-N., Liu, Q., Han, C., Zhang, X.-D., Wang, X.-Y.: Automatic extraction of imaging observation and assessment categories from breast magnetic resonance imaging reports with natural language processing. Chinese Medical Journal **132**(14), 1673–1680 (2019). doi:10.1097/CM9.0000000000000301. Accessed 2020-10-30
29. Gupta, A., Banerjee, I., Rubin, D.L.: Automatic information extraction from unstructured mammography reports using distributed semantics. Journal of Biomedical Informatics **78**, 78–86 (2018). doi:10.1016/j.jbi.2017.12.016. Accessed 2020-10-30
30. Castro, S.M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., Jacobson, R.S.: Automated annotation and classification of BI-RADS assessment from radiology reports. Journal of Biomedical Informatics **69**, 177–187 (2017). doi:10.1016/j.jbi.2017.04.011. Accessed 2020-10-30
31. Short, R.G., Bralich, J., Bogaty, D., Befera, N.T.: Comprehensive Word-Level Classification of Screening Mammography Reports Using a Neural Network Sequence Labeling Approach. Journal of Digital Imaging **32**(5), 685–692 (2019). doi:10.1007/s10278-018-0141-4. Accessed 2020-10-30
32. Lacson, R., Goodrich, M.E., Harris, K., Brawarsky, P., Haas, J.S.: Assessing Inaccuracies in Automated Information Extraction of Breast Imaging Findings. Journal of Digital Imaging **30**(2), 228–233 (2017). doi:10.1007/s10278-016-9927-4. Accessed 2020-10-30
33. Lacson, R., Harris, K., Brawarsky, P., Tosteson, T.D., Onega, T., Tosteson, A.N.A., Kaye, A., Gonzalez, I., Birdwell, R., Haas, J.S.: Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry. Journal of Digital Imaging **28**(5), 567–575 (2015). doi:10.1007/s10278-014-9762-4. Accessed 2020-10-30
34. Yim, W.-w., Kwan, S.W., Yetisgen, M.: Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. Journal of Biomedical Informatics **64**, 179–191 (2016). doi:10.1016/j.jbi.2016.10.005. Accessed 2020-10-30
35. Yim, W.-w., Kwan, S.W., Yetisgen, M.: Classifying tumor event attributes in radiology reports. Journal of the Association for Information Science and Technology **68**(11), 2662–2674 (2017). doi:10.1002/asi.23937.

Accessed 2020-10-30

36. Yim, W.-w., Denman, T., Kwan, S.W., Yetisgen, M.: Tumor information extraction in radiology reports for hepatocellular carcinoma patients. AMIA Summits on Translational Science Proceedings **2016**, 455–464 (2016). Accessed 2020-10-30
37. Pruitt, P., Naidech, A., Van Ornam, J., Borczuk, P., Thompson, W.: A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports. Emergency Radiology **26**(3), 301–306 (2019). doi:10.1007/s10140-019-01673-4. Accessed 2020-10-30
38. Farjah, F., Halgrim, S., Buist, D.S.M., Gould, M.K., Zeliadt, S.B., Loggers, E.T., Carrell, D.S.: An Automated Method for Identifying Individuals with a Lung Nodule Can Be Feasibly Implemented Across Health Systems. eGEMs **4**(1), 1254 (2016). doi:10.13063/2327-9214.1254. Accessed 2020-10-30
39. Karunakaran, B., Misra, D., Marshall, K., Mathrawala, D., Kethireddy, S.: Closing the loop — Finding lung cancer patients using NLP. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 2452–2461. IEEE, Boston, MA (2017). doi:10.1109/BigData.2017.8258203
40. Tan, W.K., Hassanpour, S., Heagerty, P.J., Rundell, S.D., Suri, P., Huhdanpaa, H.T., James, K., Carrell, D.S., Langlotz, C.P., Organ, N.L., Meier, E.N., Sherman, K.J., Kallmes, D.F., Luetmer, P.H., Griffith, B., Nerenz, D.R., Jarvik, J.G.: Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain. Academic Radiology **25**(11), 1422–1432 (2018). doi:10.1016/j.acra.2018.03.008. Accessed 2020-10-30
41. Trivedi, G., Hong, C., Dadashzadeh, E.R., Handzel, R.M., Hochheiser, H., Visweswaran, S.: Identifying incidental findings from radiology reports of trauma patients: An evaluation of automated feature representation methods. International Journal of Medical Informatics **129**, 81–87 (2019). doi:10.1016/j.ijmedinf.2019.05.021. Accessed 2020-10-30
42. Fu, S., Leung, L.Y., Wang, Y., Raulli, A.-O., Kallmes, D.F., Kinsman, K.A., Nelson, K.B., Clark, M.S., Luetmer, P.H., Kingsbury, P.R., Kent, D.M., Liu, H.: Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. JMIR Medical Informatics **7**(2), 12109 (2019). doi:10.2196/12109. Accessed 2020-10-30
43. Jnawali, K., Arbabshirani, M.R., Ulloa, A.E., Rao, N., Patel, A.A.: Automatic Classification of Radiological Report for Intracranial Hemorrhage. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC), pp. 187–190. IEEE, Newport Beach, CA, USA (2019). doi:10.1109/ICOSC.2019.8665578
44. Banerjee, I., Madhavan, S., Goldman, R.E., Rubin, D.L.: Intelligent Word Embeddings of Free-Text Radiology Reports. AMIA Annual Symposium Proceedings, 411–420 (2017). Accessed 2020-10-30
45. Kłos, M., Żyłkowski, J., Spinczyk, D.: Automatic Classification of Text Documents Presenting Radiology Examinations. In: Pietka, E., Badura, P., Kawa, J., Wieclawek, W. (eds.) Proceedings 6th International Conference Information Technology in Biomedicine(ITIB'2018). Advances in Intelligent Systems and Computing, pp. 495–505. Springer, ??? (2018). doi:10.1007/978-3-319-91211-0-43
46. Deshmukh, N., Gumustop, S., Gauriau, R., Buch, V., Wright, B., Bridge, C., Naidu, R., Andriole, K., Bizzo, B.: Semi-Supervised Natural Language Approach for Fine-Grained Classification of Medical Reports. arXiv:1910.13573 [cs.LG] (2019). Accessed 2020-10-30
47. Kim, C., Zhu, V., Obeid, J., Lenert, L.: Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLOS ONE **14**(2), 0212778 (2019). doi:10.1371/journal.pone.0212778. Accessed 2020-10-30
48. Garg, R., Oh, E., Naidech, A., Kording, K., Prabhakaran, S.: Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. Journal of Stroke and Cerebrovascular Diseases **28**(7), 2045–2051 (2019). doi:10.1016/j.jstrokecerebrovasdis.2019.02.004. Accessed 2020-10-30
49. Shin, B., Chokshi, F.H., Lee, T., Choi, J.D.: Classification of radiology reports using neural attention models. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 4363–4370. IEEE, Anchorage, AK (2017). doi:10.1109/IJCNN.2017.7966408
50. Wheater, E., Mair, G., Sudlow, C., Alex, B., Grover, C., Whiteley, W.: A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Medical Informatics and Decision Making **19**(1), 184 (2019). doi:10.1186/s12911-019-0908-7. Accessed 2020-10-30

765 51. Gorinski, P.J., Wu, H., Grover, C., Tobin, R., Talbot, C., Whalley, H., Sudlow, C., Whiteley, W., Alex, B.:
766 Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning
767 Approaches. arXiv:1903.03985 [cs.CL] (2019). Accessed 2020-10-30
768 52. Alex, B., Grover, C., Tobin, R., Sudlow, C., Mair, G., Whiteley, W.: Text mining brain imaging reports.
769 Journal of Biomedical Semantics **10**(1), 23 (2019). doi:10.1186/s13326-019-0211-7. Accessed 2020-10-30
770 53. Bozkurt, S., Gimenez, F., Burnside, E.S., Gulkesen, K.H., Rubin, D.L.: Using automatically extracted
771 information from mammography reports for decision-support. Journal of Biomedical Informatics **62**, 224–231
772 (2016). doi:10.1016/j.jbi.2016.07.001. Accessed 2020-10-30
773 54. Patel, T.A., Puppala, M., Ogunti, R.O., Ensor, J.E., He, T., Shewale, J.B., Ankerst, D.P., Kaklamani, V.G.,
774 Rodriguez, A.A., Wong, S.T.C., Chang, J.C.: Correlating mammographic and pathologic findings in clinical
775 decision support using natural language processing and data mining methods. Cancer **123**(1), 114–121 (2017).
776 doi:10.1002/cncr.30245
777 55. Banerjee, I., Bozkurt, S., Alkim, E., Sagreiya, H., Kurian, A.W., Rubin, D.L.: Automatic inference of BI-RADS
778 final assessment categories from narrative mammography report findings. Journal of Biomedical Informatics
779 **92**, 103137 (2019). doi:10.1016/j.jbi.2019.103137. Accessed 2020-10-30
780 56. Miao, S., Xu, T., Wu, Y., Xie, H., Wang, J., Jing, S., Zhang, Y., Zhang, X., Yang, Y., Zhang, X., Shan, T.,
781 Wang, L., Xu, H., Wang, S., Liu, Y.: Extraction of BI-RADS findings from breast ultrasound reports in
782 Chinese using deep learning approaches. International Journal of Medical Informatics **119**, 17–21 (2018).
783 doi:10.1016/j.ijmedinf.2018.08.009. Accessed 2020-10-30
784 57. Dunne, R.M., Ip, I.K., Abbett, S., Gershanik, E.F., Raja, A.S., Hunsaker, A., Khorasani, R.: Effect of
785 Evidence-based Clinical Decision Support on the Use and Yield of CT Pulmonary Angiographic Imaging in
786 Hospitalized Patients. Radiology **276**(1), 167–174 (2015). doi:10.1148/radiol.15141208. Accessed 2020-10-30
787 58. Banerjee, I., Ling, Y., Chen, M.C., Hasan, S.A., Langlotz, C.P., Moradzadeh, N., Chapman, B., Amrhein, T.,
788 Mong, D., Rubin, D.L., Farri, O., Lungren, M.P.: Comparative effectiveness of convolutional neural network
789 (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artificial
790 Intelligence in Medicine **97**, 79–88 (2019). doi:10.1016/j.artmed.2018.11.004. Accessed 2020-10-30
791 59. Chen, M.C., Ball, R.L., Yang, L., Moradzadeh, N., Chapman, B.E., Larson, D.B., Langlotz, C.P., Amrhein,
792 T.J., Lungren, M.P.: Deep Learning to Classify Radiology Free-Text Reports. Radiology **286**(3), 845–852
793 (2017). doi:10.1148/radiol.2017171115. Accessed 2020-10-30
794 60. Meystre, S., Gouripeddi, R., Tieder, J., Simmons, J., Srivastava, R., Shah, S.: Enhancing Comparative
795 Effectiveness Research With Automated Pediatric Pneumonia Detection in a Multi-Institutional Clinical
796 Repository: A PHIS+ Pilot Study. Journal of Medical Internet Research **19**(5), 162 (2017).
797 doi:10.2196/jmir.6887. Accessed 2020-10-30
798 61. Beyer, S.E., McKee, B.J., Regis, S.M., McKee, A.B., Flacke, S., El Saadawi, G., Wald, C.: Automatic
799 Lung-RADS™ classification with a natural language processing system. Journal of Thoracic Disease **9**(9),
800 3114–3122 (2017). doi:10.21037/jtd.2017.08.13. Accessed 2020-10-30
801 62. Patterson, O.V., Freiberg, M.S., Skanderson, M., J. Fodeh, S., Brandt, C.A., DuVall, S.L.: Unlocking
802 echocardiogram measurements for heart disease research through natural language processing. BMC
803 Cardiovascular Disorders **17**(1), 151 (2017). doi:10.1186/s12872-017-0580-8. Accessed 2020-10-30
804 63. Lee, C., Kim, Y., Kim, Y.S., Jang, J.: Automatic Disease Annotation From Radiology Reports Using Artificial
805 Intelligence Implemented by a Recurrent Neural Network. American Journal of Roentgenology **212**(4),
806 734–740 (2019). doi:10.2214/AJR.18.19869. Accessed 2020-10-30
807 64. Fiebeck, J., Laser, H., Winther, H.B., Gerbel, S.: Leaving No Stone Unturned: Using Machine Learning Based
808 Approaches for Information Extraction from Full Texts of a Research Data Warehouse. In: Auer, S., Vidal,
809 M.-E. (eds.) 13th International Conference Data Integration in the Life Sciences (DILS 2018). Lecture Notes
810 in Computer Science, pp. 50–58. Springer, Hannover, Germany (2018). doi:10.1007/978-3-030-06016-9$_5$
811 65. Hassanzadeh, H., Kholghi, M., Nguyen, A., Chu, K.: Clinical Document Classification Using Labeled and
812 Unlabeled Data Across Hospitals. AMIA Annual Symposium Proceedings **2018**, 545–554 (2018). Accessed
813 2020-10-30
814 66. Krishnan, G.S., Kamath S., S.: Ontology-driven Text Feature Modeling for Disease Prediction using
815 Unstructured Radiological Notes. Computación y Sistemas **23**(3) (2019). doi:10.13053/cys-23-3-3238.

816 Accessed 2020-10-30
817 67. Qenam, B., Kim, T.Y., Carroll, M.J., Hogarth, M.: Text Simplification Using Consumer Health Vocabulary to
818 Generate Patient-Centered Radiology Reporting: Translation and Evaluation. Journal of Medical Internet
819 Research **19**(12), 417 (2017). doi:10.2196/jmir.8536. Accessed 2020-10-30
820 68. Lafourcade, M., Ramadier, L.: Radiological text simplification using a general knowledge base. In: 18th
821 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017).
822 CICLing 2017. Budapest, Hungary, ??? (2017). doi:10.1007/978-3-319-77116-8_46
823 69. Hong, Y., Zhang, J.: Investigation of Terminology Coverage in Radiology Reporting Templates and Free-text
824 Reports. International Journal of Knowledge Content Development & Technology **5**, 5–14 (2015).
825 doi:10.5865/IJKCT.2015.5.1.005
826 70. Comelli, A., Agnello, L., Vitabile, S.: An ontology-based retrieval system for mammographic reports. In: 2015
827 IEEE Symposium on Computers and Communication (ISCC), pp. 1001–1006. IEEE, Larnaca (2015).
828 doi:10.1109/ISCC.2015.7405644
829 71. Cotik, V., Filippo, D., Castano, J.: An Approach for Automatic Classification of Radiology Reports in Spanish.
830 Studies in Health Technology and Informatics **216**, 634–638 (2015)
831 72. Johnson, E., Baughman, W.C., Ozsoyoglu, G.: A method for imputation of semantic class in diagnostic
832 radiology text. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.
833 750–755. IEEE, Washington, DC (2015). doi:10.1109/BIBM.2015.7359780
834 73. Mujjiga, S., Krishna, V., Chakravarthi, K., J, V.: Identifying Semantics in Clinical Reports Using Neural
835 Machine Translation. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 9552–9557
836 (2019). doi:10.1609/aaai.v33i01.33019552. Accessed 2020-10-30
837 74. Lafourcade, M., Ramadier, L.: Semantic RelationExtraction with Semantic Patterns: Experiment on Radiology
838 Report. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC
839 2016). LREC 2016 Proceedings. European Language Resources Association (ELRA), Portorož, Slovenia
840 (2016). https://hal.archives-ouvertes.fr/hal-01382320
841 75. Shelmerdine, S.C., Singh, M., Norman, W., Jones, R., Sebire, N.J., Arthurs, O.J.: Automated data extraction
842 and report analysis in computer-aided radiology audit: practice implications from post-mortem paediatric
843 imaging. Clinical Radiology **74**(9), 733–1173318 (2019). doi:10.1016/j.crad.2019.04.021. Accessed 2020-10-30
844
845 76. Mabotuwana, T., Hombal, V., Dalal, S., Hall, C.S., Gunn, M.: Determining Adherence to Follow-up Imaging
846 Recommendations. Journal of the American College of Radiology **15**(3, Part A), 422–428 (2018).
847 doi:10.1016/j.jacr.2017.11.022. Accessed 2020-10-30
848 77. Dalal, S., Hombal, V., Weng, W.-H., Mankovich, G., Mabotuwana, T., Hall, C.S., Fuller, J., Lehnert, B.E.,
849 Gunn, M.L.: Determining Follow-Up Imaging Study Using Radiology Reports. Journal of Digital Imaging
850 **33**(1), 121–130 (2020). doi:10.1007/s10278-019-00260-w. Accessed 2020-10-30
851 78. Bobbin, M.D., Ip, I.K., Sahni, V.A., Shinagare, A.B., Khorasani, R.: Focal Cystic Pancreatic Lesion Follow-up
852 Recommendations After Publication of ACR White Paper on Managing Incidental Findings. Journal of the
853 American College of Radiology **14**(6), 757–764 (2017). doi:10.1016/j.jacr.2017.01.044. Accessed 2020-10-30
854 79. Kwan, J.L., Yermak, D., Markell, L., Paul, N.S., Shojania, K.J., Cram, P.: Follow Up of Incidental High-Risk
855 Pulmonary Nodules on Computed Tomography Pulmonary Angiography at Care Transitions. Journal of
856 Hospital Medicine **14**(6), 349–352 (2019). doi:10.12788/jhm.3128. Accessed 2020-10-30
857 80. Mabotuwana, T., Hall, C.S., Tieder, J., Gunn, M.L.: Improving Quality of Follow-Up Imaging
858 Recommendations in Radiology. AMIA Annual Symposium Proceedings **2017**, 1196–1204 (2018). Accessed
859 2020-10-30
860 81. Brown, A.D., Marotta, T.R.: A Natural Language Processing-based Model to Automate MRI Brain Protocol
861 Selection and Prioritization. Academic Radiology **24**(2), 160–166 (2017). doi:10.1016/j.acra.2016.09.013.
862 Accessed 2020-10-30
863 82. Trivedi, H., Mesterhazy, J., Laguna, B., Vu, T., Sohn, J.H.: Automatic Determination of the Need for
864 Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing
865 Algorithm. Journal of Digital Imaging **31**(2), 245–251 (2018). doi:10.1007/s10278-017-0021-3. Accessed
866 2020-10-30

83. Zhang, A.Y., Lam, S.S.W., Liu, N., Pang, Y., Chan, L.L., Tang, P.H.: Development of a Radiology Decision Support System for the Classification of MRI Brain Scans. In: 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), pp. 107–115 (2018). doi:10.1109/BDCAT.2018.00021
84. Brown, A.D., Marotta, T.R.: Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. Journal of the American Medical Informatics Association **25**(5), 568–571 (2018). doi:10.1093/jamia/ocx125. Accessed 2020-10-30
85. Yan, Z., Ip, I.K., Raja, A.S., Gupta, A., Kosowsky, J.M., Khorasani, R.: Yield of CT Pulmonary Angiography in the Emergency Department When Providers Override Evidence-based Clinical Decision Support. Radiology **282**(3), 717–725 (2016). doi:10.1148/radiol.2016151985. Accessed 2020-10-30
86. Kang, S.K., Garry, K., Chung, R., Moore, W.H., Iturrate, E., Swartz, J.L., Kim, D.C., Horwitz, L.I., Blecker, S.: Natural Language Processing for Identification of Incidental Pulmonary Nodules in Radiology Reports. Journal of the American College of Radiology **16**(11), 1587–1594 (2019). doi:10.1016/j.jacr.2019.04.026. Accessed 2020-10-30
87. Brown, A.D., Kachura, J.R.: Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization. Journal of the American College of Radiology **16**(6), 840–844 (2019). doi:10.1016/j.jacr.2018.12.004. Accessed 2020-10-30
88. Grundmeier, R.W., Masino, A.J., Casper, T.C., Dean, J.M., Bell, J., Enriquez, R., Deakyne, S., Chamberlain, J.M., Alpern, E.R.: Identification of Long Bone Fractures in Radiology Reports Using Natural Language Processing to Support Healthcare Quality Improvement. Applied Clinical Informatics **7**(4), 1051–1068 (2016). doi:10.4338/ACI-2016-08-RA-0129. Accessed 2020-10-30
89. Heilbrun, M.E., Chapman, B.E., Narasimhan, E., Patel, N., Mowery, D.: Feasibility of Natural Language Processing–Assisted Auditing of Critical Findings in Chest Radiology. Journal of the American College of Radiology **16**(9, Part B), 1299–1304 (2019). doi:10.1016/j.jacr.2019.05.038. Accessed 2020-10-30
90. Maros, M.E., Wenz, R., Förster, A., Froelich, M.F., Groden, C., Sommer, W.H., Schönberg, S.O., Henzler, T., Wenz, H.: Objective Comparison Using Guideline-based Query of Conventional Radiological Reports and Structured Reports. In Vivo **32**(4), 843–849 (2018). doi:10.21873/invivo.11318. Accessed 2020-10-30
91. Minn, M.J., Zandieh, A.R., Filice, R.W.: Improving Radiology Report Quality by Rapidly Notifying Radiologist of Report Errors. Journal of Digital Imaging **28**(4), 492–498 (2015). doi:10.1007/s10278-015-9781-9. Accessed 2020-10-30
92. Goldshtein, I., Chodick, G., Kochba, I., Gal, N., Webb, M., Shibolet, O.: Identification and Characterization of Nonalcoholic Fatty Liver Disease. Clinical Gastroenterology and Hepatology **18**(8), 1887–1889 (2020). doi:10.1016/j.cgh.2019.08.007. Accessed 2020-10-30
93. Redman, J.S., Natarajan, Y., Hou, J.K., Wang, J., Hanif, M., Feng, H., Kramer, J.R., Desiderio, R., Xu, H., El-Serag, H.B., Kanwal, F.: Accurate Identification of Fatty Liver Disease in Data Warehouse Utilizing Natural Language Processing. Digestive Diseases and Sciences **62**(10), 2713–2718 (2017). doi:10.1007/s10620-017-4721-9. Accessed 2020-10-30
94. Sada, Y., Hou, J., Richardson, P., El-Serag, H., Davila, J.: Validation of Case Finding Algorithms for Hepatocellular Cancer from Administrative Data and Electronic Health Records using Natural Language Processing. Medical care **54**(2), 9–14 (2016). doi:10.1097/MLR.0b013e3182a30373. Accessed 2020-10-30
95. Li, A.Y., Elliot, N.: Natural language processing to identify ureteric stones in radiology reports. Journal of Medical Imaging and Radiation Oncology **63**(3), 307–310 (2019). doi:10.1111/1754-9485.12861. Accessed 2020-10-30
96. Tan, W.K., Heagerty, P.J.: Surrogate-guided sampling designs for classification of rare outcomes from electronic medical records data. arXiv:1904.00412 [stat.ME] (2019). Accessed 2020-10-30
97. Yadav, K., Sarioglu, E., Choi, H.-A., Cartwright, W.B., Hinds, P.S., Chamberlain, J.M.: Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. Academic Emergency Medicine **23**(2), 171–178 (2016). doi:10.1111/acem.12859. Accessed 2020-10-30
98. Mahan, M., Rafter, D., Casey, H., Engelking, M., Abdallah, T., Truwit, C., Oswood, M., Samadani, U.: tbiExtractor: A framework for extracting traumatic brain injury common data elements from radiology reports. bioRxiv 585331 (2019). doi:10.1101/585331. Accessed 2020-12-05

99. Brizzi, K., Zupanc, S.N., Udelsman, B.V., Tulsky, J.A., Wright, A.A., Poort, H., Lindvall, C.: Natural Language Processing to Assess Palliative Care and End-of-Life Process Measures in Patients With Breast Cancer With Leptomeningeal Disease. American Journal of Hospice and Palliative Medicine **37**(5), 371–376 (2019). doi:10.1177/1049909119885585. Accessed 2020-10-30
100. Van Haren, R.M., Correa, A.M., Sepesi, B., Rice, D.C., Hofstetter, W.L., Mehran, R.J., Vaporciyan, A.A., Walsh, G.L., Roth, J.A., Swisher, S.G., Antonoff, M.B.: Ground Glass Lesions on Chest Imaging: Evaluation of Reported Incidence in Cancer Patients Using Natural Language Processing. The Annals of Thoracic Surgery **107**(3), 936–940 (2019). doi:10.1016/j.athoracsur.2018.09.016. Accessed 2020-10-30
101. Noorbakhsh-Sabet, N., Tsivgoulis, G., Shahjouei, S., Hu, Y., Goyal, N., Alexandrov, A.V., Zand, R.: Racial Difference in Cerebral Microbleed Burden Among a Patient Population in the Mid-South United States. Journal of Stroke and Cerebrovascular Diseases **27**(10), 2657–2661 (2018). doi:10.1016/j.jstrokecerebrovasdis.2018.05.031. Accessed 2020-10-30
102. Gould, M.K., Tang, T., Liu, I.-L.A., Lee, J., Zheng, C., Danforth, K.N., Kosco, A.E., Di Fiore, J.L., Suh, D.E.: Recent Trends in the Identification of Incidental Pulmonary Nodules. American Journal of Respiratory and Critical Care Medicine **192**(10), 1208–1214 (2015). doi:10.1164/rccm.201505-0990OC. Accessed 2020-10-30
103. Huhdanpaa, H.T., Tan, W.K., Rundell, S.D., Suri, P., Chokshi, F.H., Comstock, B.A., Heagerty, P.J., James, K.T., Avins, A.L., Nedeljkovic, S.S., Nerenz, D.R., Kallmes, D.F., Luetmer, P.H., Sherman, K.J., Organ, N.L., Griffith, B., Langlotz, C.P., Carrell, D., Hassanpour, S., Jarvik, J.G.: Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. Journal of Digital Imaging **31**(1), 84–90 (2018). doi:10.1007/s10278-017-0013-3. Accessed 2020-10-30
104. Masino, A.J., Grundmeier, R.W., Pennington, J.W., Germiller, J.A., Crenshaw, E.B.: Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. BMC Medical Informatics and Decision Making **16**(1), 65 (2016). doi:10.1186/s12911-016-0306-3. Accessed 2020-10-30
105. Valtchinov, V.I., Lacson, R., Wang, A., Khorasani, R.: Comparing Artificial Intelligence Approaches to Retrieve Clinical Reports Documenting Implantable Devices Posing MRI Safety Risks. Journal of the American College of Radiology **17**(2), 272–279 (2020). doi:10.1016/j.jacr.2019.07.018. Accessed 2020-10-30
106. Zech, J., Forde, J., Titano, J.J., Kaji, D., Costa, A., Oermann, E.K.: Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Annals of Translational Medicine **7**(11) (2019). doi:10.21037/atm.2018.08.11. Accessed 2020-10-30
107. Zhang, Y., Merck, D., Tsai, E.B., Manning, C.D., Langlotz, C.P.: Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. arXiv:1911.02541 [cs.CL] (2019). Accessed 2020-10-30
108. Steinkamp, J.M., Chambers, C., Lalevic, D., Zafar, H.M., Cook, T.S.: Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. Journal of Digital Imaging **32**(4), 554–564 (2019). doi:10.1007/s10278-019-00234-y. Accessed 2020-10-30
109. Cocos, A., Qian, T., Callison-Burch, C., Masino, A.J.: Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. Journal of Biomedical Informatics **69**, 86–92 (2017). doi:10.1016/j.jbi.2017.04.003. Accessed 2020-10-30
110. Ratner, A., Hancock, B., Dunnmon, J., Goldman, R., Ré, C.: Snorkel MeTaL: Weak Supervision for Multi-Task Learning. In: Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning. DEEM'18, vol. 3, pp. 1–4. ACM, Houston, TX, USA (2018). doi:10.1145/3209889.3209898. https://doi.org/10.1145/3209889.3209898 Accessed 2020-10-30
111. Zhu, H., Paschalidis, I.C., Hall, C., Tahmasebi, A.: Context-Driven Concept Annotation in Radiology Reports: Anatomical Phrase Labeling. AMIA Summits on Translational Science Proceedings **2019**, 232–241 (2019). Accessed 2020-10-30
112. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, (2013). http://arxiv.org/abs/1301.3781
113. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
114. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in Pre-Training Distributed Word Representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)

115. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018). _eprint: 1802.05365
116. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
117. National Library of Medicine: Unified Medical Language System, (2021). https://www.nlm.nih.gov/research/umls/index.html
118. RSNA: RadLex, (2021). http://radlex.org/
119. National Library of Medicine: SNOMED CT, (2021). @miscumls, author = National Library of Medicine, title = SNOMED CT, url = https://www.nlm.nih.gov/healthit/snomedct/index.html, year = 2021, urldate=Accessed 07 Feb 2021
120. Bulu, H., Sippo, D.A., Lee, J.M., Burnside, E.S., Rubin, D.L.: Proposing New RadLex Terms by Analyzing Free-Text Mammography Reports. Journal of Digital Imaging **31**(5), 596–603 (2018). doi:10.1007/s10278-018-0064-0. Accessed 2020-10-30
121. Hassanpour, S., Langlotz, C.P.: Unsupervised Topic Modeling in a Large Free Text Radiology Report Repository. Journal of Digital Imaging **29**(1), 59–62 (2016). doi:10.1007/s10278-015-9823-3. Accessed 2020-10-30
122. Zhao, Y., Fesharaki, N.J., Liu, H., Luo, J.: Using data-driven sublanguage pattern mining to induce knowledge models: application in medical image reports knowledge representation. BMC Medical Informatics and Decision Making **18**(1), 61 (2018). doi:10.1186/s12911-018-0645-3. Accessed 2020-10-30
123. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1), 37–46 (1960). doi:10.1177/001316446002000104. Accessed 2020-10-31
124. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics **22**(5), 1589–1604 (2018). doi:10.1109/JBHI.2017.2767063
125. Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storlie, C.B., Habermann, E.B., Naessens, J.M., Larson, D.W., Liu, H.: Deep learning and alternative learning strategies for retrospective real-world clinical data. npj Digital Medicine **2**(1), 1–5 (2019). doi:10.1038/s41746-019-0122-0. Accessed 2020-12-02
126. Yang, H., Li, L., Yang, R., Zhou, Y.: Towards Automated Knowledge Discovery of Hepatocellular Carcinoma: Extract Patient Information from Chinese Clinical Reports. In: Proceedings of the 2nd International Conference on Medical and Health Informatics. ICMHI '18, pp. 111–116. ACM, New York, NY, USA (2018). doi:10.1145/3239438.3239445. https://doi.org/10.1145/3239438.3239445 Accessed 2020-10-30
127. Wood, D.A., Lynch, J., Kafiabadi, S., Guilhem, E., Busaidi, A.A., Montvila, A., Varsavsky, T., Siddiqui, J., Gadapa, N., Townend, M., Kiik, M., Patel, K., Barker, G., Ourselin, S., Cole, J.H., Booth, T.C.: Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). arXiv:2002.06588 [cs.CV] (2020). Accessed 2020-12-03
128. Ong, C.J., Orfanoudaki, A., Zhang, R., Caprasse, F.P.M., Hutch, M., Ma, L., Fard, D., Balogun, O., Miller, M.I., Minnig, M., Saglam, H., Prescott, B., Greer, D.M., Smirnakis, S., Bertsimas, D.: Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLOS ONE **15**(6), 0234908 (2020). doi:10.1371/journal.pone.0234908. Accessed 2020-10-31
129. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1500–1519. Association for Computational Linguistics, Online (2020). doi:10.18653/v1/2020.emnlp-main.117. https://www.aclweb.org/anthology/2020.emnlp-main.117 Accessed 2020-12-03
130. Grivas, A., Alex, B., Grover, C., Tobin, R., Whiteley, W.: Not a cute stroke: Analysis of Rule- and Neural Network-Based Information Extraction Systems for Brain Radiology Reports. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (2020)
131. Ettinger, A.: What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. Transactions of the Association for Computational Linguistics **8**, 34–48 (2020). doi:10.1162/tacl_a_00298. Accessed 2020-10-31
132. Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., McDermott, M.: Publicly

Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). doi:10.18653/v1/W19-1909. https://www.aclweb.org/anthology/W19-1909

133. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. CoRR **abs/2004.09167** (2020). _eprint: 2004.09167
134. Yasaka, K., Abe, O.: Deep learning and artificial intelligence in radiology: Current applications and future directions. PLOS Medicine **15**(11), 1002707 (2018). doi:10.1371/journal.pmed.1002707. Accessed 2020-11-01
135. Percha, B., Zhang, Y., Bozkurt, S., Rubin, D., Altman, R.B., Langlotz, C.P.: Expanding a radiology lexicon using contextual patterns in radiology reports. Journal of the American Medical Informatics Association **25**(6), 679–685 (2018). doi:10.1093/jamia/ocx152. Accessed 2020-10-30
136. Tahmasebi, A.M., Zhu, H., Mankovich, G., Prinsen, P., Klassen, P., Pilato, S., van Ommering, R., Patel, P., Gunn, M.L., Chang, P.: Automatic Normalization of Anatomical Phrases in Radiology Reports Using Unsupervised Learning. Journal of Digital Imaging **32**(1), 6–18 (2019). doi:10.1007/s10278-018-0116-5. Accessed 2020-10-30
137. Banerjee, I., Chen, M.C., Lungren, M.P., Rubin, D.L.: Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. Journal of Biomedical Informatics **77**, 11–20 (2018). doi:10.1016/j.jbi.2017.11.012. Accessed 2020-10-30
138. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine **13**(3), 55–75 (2018). doi:10.1109/MCI.2018.2840738
139. Donnelly, L.F., Grzeszczuk, R., Guimaraes, C.V., Zhang, W., Bisset III, G.S.: Using a Natural Language Processing and Machine Learning Algorithm Program to Analyze Inter-Radiologist Report Style Variation and Compare Variation Between Radiologists When Using Highly Structured Versus More Free Text Reporting. Current Problems in Diagnostic Radiology **48**(6), 524–530 (2019). doi:10.1067/j.cpradiol.2018.09.005. Accessed 2020-10-30
140. Xie, Z., Yang, Y., Wang, M., Li, M., Huang, H., Zheng, D., Shu, R., Ling, T.: Introducing Information Extraction to Radiology Information Systems to Improve the Efficiency on Reading Reports. Methods of Information in Medicine **58**(2-03), 94–106 (2019). doi:10.1055/s-0039-1694992
141. Zech, J., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., Costa, A., Bederson, J., Lehar, J., Oermann, E.K.: Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports. Radiology **287**(2), 570–580 (2018). doi:10.1148/radiol.2018171093. Accessed 2020-10-30
142. Yim, W.-w., Kwan, S.W., Johnson, G., Yetisgen, M.: Classification of hepatocellular carcinoma stages from free-text clinical and radiology reports. AMIA Annual Symposium Proceedings **2017**, 1858–1867 (2018). Accessed 2020-10-30
143. Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., , He, Z.: Explainable artificial intelligence models using real-world electronic health records data: a systematic scoping review. Journal of the American Medical Informatics Association (2020)
144. Dong, H., Suárez-Paniagua, V., Whiteley, W., Wu, H.: Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. Journal of Biomedical Informatics **116**, 103728 (2021). doi:10.1016/j.jbi.2021.103728

Figure Legends

Figure 1 - PRISMA diagram for search publication retrieval

Figure 2 Clinical application of publication by year

Figure 3 - NLP method breakdown

Figure 4 - NLP method by year

Figure 5 - Application Category and NLP Method, Mean and Median Summaries. Mean value is indicated by a vertical bar, the box shows error bars and the asterisk is the median value.

Table Legends

Table 1 - Metadata enriching steps undertaken for each publication

Table 2 - Automated filtering steps to remove irrelevant publications

Table 3 - Scan modality

Table 4 - Image sampling method

Table 5 - Anatomical region scanned

Table 6 - Disease category

Table 7 - Radiology report language

Table 8 - Clinical application category by technical objective

Table 9 - Breakdown of NLP method

Table 10 - NLP Method by data size properties, minimum data size, maximum data size and median value, studies reporting in numbers of radiology reports

Table 11 - Grouped data size and number of studies in each group, only for studies reporting in numbers of radiology reports

Table 12 - Studies reporting on total data size used and details on training set size, validation set size, test set size and annotation set size

Additional Files

Additional file 1 — Publications Reviewed Table

File contains a table which categorises all 164 publications analysed into clinical application area and anatomical region (where applicable). File is a Word document.

Additional file 2 — Individual publication data

File contains one row per publications for all 164 publications detailing: clinical category, technical category, image sampling method, language of reports, anatomical region, imaging modality, disease area, data type, total data size, annotation set size, training size, validation set size,test set size. File is a CSV file.